異常値の検出

野口和也

本論は回帰分析の場における異常値（集団）の検出について論ずるものである。元来，異常値に対する処理は，それを見つけてデータより除去してしまうか，前もって異常値の存在を見込んだ上で robust な推定方式を採用するかのいずれかである。後者は何らかの理由により異常値の検出がなし得ない場合にのみ行うべき方法であるが，社会科学における統計データはその生成メカニズムが複雑であることが多く，異常値の検出は決して易しくないからある程度やむを得ない。さらに悪いことには，社会科学のデータには異常値が集団として存在している場合がしばしばあり，さらに問題を困難なものとしている。しかし，データの異常性について何らの検討もおこなわずに徒に robust 推定をふりまわすのも決して賢明なやり方とは思えない。本論では，この異常値の検出の問題を，回帰分析を念頭において単一異常値と複数個の異常値の場合に分けて検討し，さらにそれらが検出されたときの処理方法についても論及する。

1. 異常値の定義

「異常値（outlier）」という言葉は統計分析ではよく用いられるが，これに定義らしい定義は与えられていない。よくあるのは‘他の観測値と著しくかけ離れた観測値’というような表現であるが，もちろんこれでは定義にならい。率直に言って，異常値とはその定義を与えにくいもののひとつである。なんとなれば，分析目的によって，ある観測値が異常値であったりなかったりすることが大いに起こり得るからである。本論は分析目的＝回帰分析と想定し，「回帰における異常値の検出」について論ずるものである。通常それは‘残差が他の観測値と比較して異常に大きい観測値（集団）’という様に表現されているが，以下では，‘他の観測値に比べて回帰の諸結果（母数・従属変数の推定値，残差，およびそれらの分散・共分散）に与える影響力が大である観測値’，すなわ

ち回帰に influentiul な観測値を異常値として考えることにする。

2. 異常値の源泉と処理

異常値の原因を次の四点に分類して考えてみよう。実際には **b**) と [**c**), **d**)] の区別は難しい。

a) 観測誤差および集計誤差

これはデータそのものの記録ミスや，データをコンピュータ用に変換する際のミスに代表される。このような種類の誤差は通常ケタ誤りとして起こることが多いため異常値として明確にとらえやすいと言える。異常値が検出され，その原因がこれらに因るものであることが明らかであれば，データから除外することが賢明である。例をあげると，数字の1と7は書き誤りが多い。しかし100.1を700.1としたとのミスは把えやすいが，100.1を100.7としたときのミスは永久に発見できない可能性があることをつけ加えておこう。

b) モデルの mis-specification

モデルの想定が誤っているために異常値らしきものが発生している場合である。データの生成メカニズムが想定されたモデルによってうまく説明されていないのであるから，これは厳密な意味での異常値ではない。しかし“真の”モデルが完全に把握できない限り（この場合が普通），モデルの想定が悪くて異常値に見えるのか，本来の意味で異常値なのかを認定することはできない。例えば，真のデータ生成メカニズムにとって重要な役割を演ずる変数がモデルより欠落している場合や，何らかの質的・制度的要因の変化がデータの一部分に作用しているような場合である。もし想定ミスによる異常値であることがはっきりしているならば，それをデータより除外することは賢明でない。そのような場合には‘異常値’は母集団に関する有益な情報を提供してくれているのであるから，十分にこれを利用しモデルの想定に還元すべきである。

c) 誤差分布の異常性

これは誤差の型の想定に関して起こる問題である。実際には殆んど正規性の仮定である。例えば Cauchy 分布のような極端に裾の広い分布 (heavily-tailed distribution) に誤差項が従う場合，正規性の仮定のもとでは説明することのできない観測値が発生しやすい。この場合，分布型の想定を変えればよいのであるが，実際問題として正規分布以外の仮定を用いると分析が厄介なものになるのでデータそのものを変換してしまうか，説明不可能な観測値を削除してしまうことが多い。

d) Dixon mechanism

最後に，「混合分布 (contaminating distribution)」の問題について触れておかねばならない。これは複数個の異常値の問題に密接に関連しており，現在多方面からの研究が行なわれているテーマである[1]。Dixon (1953) はサイズ n の標本において $n-k$ 個の観測値が 'basic distribution' $f_0(x)$ より生じ，残りの k 個の観測値が混合分布 $f_1(x)$ より生じているという mechanism を考えた。この mechanism をモデル化したものを「混合モデル (mixture model)」と呼び，

$$(2.1)\quad F(x)=\left(1-\frac{k}{n}\right)f_0(x)+\frac{k}{n}f_1(x)$$

で表わす。k/n はあるひとつの観測値が混合分布より生じている確率であるが，通常未知である。もし k/n が既知で，$f_0(x)=f_0(x+\lambda)$ ならば (λ は定数)，

$$(2.2)\quad x_j(j=1,2,\cdots\cdots,n-k)\sim f_0(x),\ x_j(j=k,k+1,\cdots\cdots,n)\sim f_0(x+\lambda)$$

となり，これは同型分布の位置をずらしたものであるところから「slippage model」と呼ばれる[2]。これらのモデルの考察は本論のテーマからはずれるためこれ以上続けるわけにはいかないので一例を挙げるにとどめたい。図1を見ていただきたい。もし

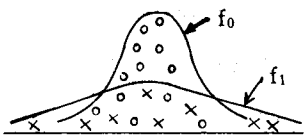

図 1

f_0 と f_1 が図のような重なり方をしていたらどうであろうか。○印が f_0, ×印が f_1 より生じた観測値である。このような場合，○も×もうまくおさまるような，f_0 でも f_1 でもない分布から全ての観測値が生成されている，という結論に陥る可能性は大きい。

注(1) 最近の文献では Mosteller = Tukey (1977), Gnanadesikan (1977), Barnett = Lewis (1978), Hawkins (1980) が挙げられる (paper を除いた)。これらは単純な度数分布概念の拡張から，多変量解析，non-parametrics に至るまで多様な角度からのアプローチとなっている。

(2) (2.2) を Dixon model と呼ぶこともある。(2.1) および (2.2) については David (1979) 61-73 p. を参照されたい。

3. 異常値の検出

p 個の独立変数を含む母集団回帰 $\boldsymbol{y}=\boldsymbol{x}\beta+\epsilon$ を $\boldsymbol{y}=\boldsymbol{xb}+\boldsymbol{e}$ で推定する問題(1)を考えよう。回帰分析の場における異常値というのは，回帰の結果として得られる諸々の統計量に対して，他の観測値と比較して極端に大きな影響力を持つ観測値（集団）である。以下では影響力の強さを検討するために，個々の観測値はデータ行列 $\boldsymbol{x}$ のひとつの列とそれに対応する $\boldsymbol{y}$ の要素に直接的に関連しているという点に注目して，異常値とおぼしき観測値を除去したときに諸々の統計量にどのような変化が現われるかを考察する。

a) 基本統計量 h

h_i を次式で定義される $n\times n$ 行列（n は観測値数）の対角要素であるとしよう(2)。

$$(3.1)\quad \boldsymbol{H}=\boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}' \rightleftarrows h_i=\boldsymbol{x}_i^r(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}_i^{r\prime}$$

$\hat{\boldsymbol{y}}\equiv\boldsymbol{xb}=\boldsymbol{Hy}$ から h_i には推定値 $\hat{y}_i$ に関する情報がおさめられていることがわかる。変数の中心化を ～ をつけて表わすことにすると，

$$(3.2)\quad \tilde{h}_i=(\boldsymbol{x}_i^r-\bar{\boldsymbol{x}}^r)(\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}')^{-1}(\boldsymbol{x}_i^r-\bar{\boldsymbol{x}}^r)'$$

となる[3]から，$\tilde{h}_i$ は距離を表わす測度とも考えることができる。

$\boldsymbol{H}$ の性質から，$0 \leqslant h_i \leqslant 1$ であり，$\boldsymbol{x}$ がフルランクであることから，

$$\text{(3. 3)} \quad \sum_{i=1}^{n} h_i = p$$

である。したがって h_i の平均的な大きさは p/n であり，これより大きな値をもたらす観測値ほど異常値らしく見えることになる。個々の観測値がほぼ p/n に近い h_i をもつようなデータは，この意味で望ましいデータである。h_i の限界点（cutoff point：これ以上大きな h_i では観測値 i を異常値と見なす点）は独立変数が独立に分布しているときにのみ得られる。Rao (1973) は，多変量正規モデルの仮定下では

$$\text{(3. 4)} \quad \frac{n-p}{p-1} \frac{\left[h_i - \left(\frac{1}{n}\right)\right]}{1-h_i} \sim F_{p-1,\ n-p}$$

となることを示している[4]。$p>10$ および $n-p>50$ に対しては F 値の 95% 点は 2 以下であり，$p/n>0.4$ では全ての観測値は疑しくなる。データごとに (3.4) 式の統計量を計算してもよいが，通常は平均的な大きさの倍数である $2p/n$ を限界点とすれば十分であろう。

b) $\boldsymbol{e}$

回帰残差を用いて Gauss-Markoff 型のモデルの仮定を検討する方法は，これまで伝統的に用いられてきたものである[5]。加えて誤差項の正規性に関するテストを行なうことも可能で，最小二乗推定量が誤差分布の正規性に関係なく BLUE の性質を持つことから計量経済学的な問題ではこのようなテストは見落されがちであるが，推定効率や検出力を考えれば（これらを考えないような分析はそれ自体好ましいものではない）統計分析において不可欠な一手順である。誤差項分布の有害な正規性からの乖離は，skewness, 多峰分布性，thick-tailed distribution, 等によって代表される。

利用されることの多いグラフィックな手法[6]は検出力を犠牲にしてその簡便

性を強調したものである。Gauss-Markoff 型モデルの諸仮定の検討には残差の度数分布やその累積分布を用いればよい。正規性の検討には正規確率プロットを用いればよく，これは累積正規分布をその勾配が標準偏差で，その切片が平均値であるような直線として表わすものである。残差が正規分布をしていないようであれば累積残差のプロットが直線からの乖離となって表われ，異常値は累積分布直線の両端に現われる。

残差を用いて異常値検出のための測度（統計量）を考えてみよう。上述のグラフィックな手法では e_i を ϵ_i の推定値として用いるわけであるが，$\mathrm{var}(e_i)=\sigma^2(1-h_i)$ である[7]ため，$p/n \gg 0$ でない限り，Gauss-Markoff の仮定のもとでも e_i の分散の大きさは一定でない。この欠点を補うために，「標準化残差 (standardized residual)」

$$(3.5)\quad e_{si}\equiv\frac{e_i}{s\sqrt{1-h_i}}$$

が考えられている。ここでは i 番目の観測値の除去の効果を検討するのが目的であるから，

$$(3.6)\quad e_i^*\equiv\frac{e_i}{s(i)\sqrt{1-h_i}}$$

を用いることにする。$s(i)$ は $\boldsymbol{x}_i^r$（と y_i）を除去した場合の誤差分散の推定量 $s^2(i)$ の平方根である。(3.6) は「スチューデント化された残差 (studentized residual)」と呼ばれ，Lund (1975) が有意水準 0.10, 0.05, 0.01 に対する限界点表 ($5\leqq n\leqq 100$, $1\leqq p\leqq 25$) を作成している[8]。Lund 表によらなくても，e_i^* は $n-p-1$ を自由度とする t 分布に近似的に従うことを利用すればよい。

c) $\boldsymbol{b}$

観測値 i の除去が回帰母数の推定量ベクトル $\boldsymbol{b}$ に与える作用の大きさは，$\boldsymbol{b}(i)$ で $\boldsymbol{x}$ および $\boldsymbol{y}$ から i 列を除去した状態の推定量ベクトルを表わすことにすると，

(3. 7) $\Delta \boldsymbol{b}(i) \equiv \boldsymbol{b} - \boldsymbol{b}(i) = \dfrac{(\boldsymbol{x}'\boldsymbol{x})^{-1}(\boldsymbol{x}_i{}^r)' e_i}{1-h_i}$

として得られる。次に，個々の推定量 b_j が受ける作用の大きさを調べるため，$\Delta \boldsymbol{b}(i)$ を b_j の分散 $\sigma^2(\boldsymbol{x}'\boldsymbol{x})_{jj}^{-1}$ によりスケーリングしてみよう。$\boldsymbol{c} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'$ とすれば，

(3. 8) $b_j - b_j(i) = \dfrac{c_{ji} e_i}{1-h_i}$

であり，

(3. 9) $\sum_{i=1}^{n} (\boldsymbol{x}'\boldsymbol{x})^{-1}(\boldsymbol{x}_i{}^r)' \boldsymbol{x}_i{}^r (\boldsymbol{x}'\boldsymbol{x})^{-1} = (\boldsymbol{x}'\boldsymbol{x})^{-1}$

であることから

(3. 10) $\mathrm{var}(b_j) = \sigma^2 \sum_{k=1}^{n} c_{jk}{}^2$

を得る。したがって，観測値 i を除去することにより推定量 b_j が受ける影響の大きさの標準測度

(3. 11) $\Delta^* b_j(i) \equiv \dfrac{b_j - b_j(i)}{s(i)\sqrt{(\boldsymbol{x}'\boldsymbol{x})_{jj}^{-1}}} = \dfrac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}{}^2}} \dfrac{e_i}{s(i)(1-h_i)}$

が定義される。ここで，$s^2(i) = \dfrac{1}{n-p-1} \sum_{k \neq 1} [y_k - \boldsymbol{x}_k{}^r \boldsymbol{b}(i)]^2$ であり，$s(i)$ の関係式

(3. 12) $(n-p-1)s^2(i) = (n-p)s^2 - \dfrac{e_j{}^2}{1-h_i}$

を利用した。$\Delta^* b_j(i)$ の絶対値の大きい値は b_j に対して influential な力をもつ。

d) $\hat{y}_i$

$\boldsymbol{b}$ の場合と同じ考え方で，従属変数の推定量[9] $\hat{\boldsymbol{y}}$ に対する観測値 i の除去の効果を検討することができる。$\hat{y}(i)$ で $\boldsymbol{x}$ および $\boldsymbol{y}$ から i 列を除去した状態の推定量ベクトルを表わすと，

(3. 13) $\Delta\hat{y}(i)\equiv\hat{y}_i-\hat{y}_i(i)=\boldsymbol{x}_i'[\boldsymbol{b}-\boldsymbol{b}(i)]=\dfrac{h_ie_i}{1-h_i}$

を得る。スケーリングのため，$\hat{y}_i=\boldsymbol{x}_i'\boldsymbol{b}$ の標準偏差 $\sigma\sqrt{h_i}$ の逆数をかければ，

(3. 14) $\Delta^*\hat{y}(i)\equiv\left[\dfrac{h_i}{1-h_i}\right]^{\frac{1}{2}}\dfrac{e_i}{s(i)\sqrt{1-h_i}}$

となり，観測値 i を除去したときの推定量への影響を表わす測度となる（σ は $s(i)$ によって推定されている）。

e) Cov($\boldsymbol{b}$)

全ての観測値よりもたされる $\boldsymbol{b}$ の共分散行列 $\sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1}$ と，観測値 i が除去された状態での $\boldsymbol{b}(i)$ の共分散行列 $\sigma^2[\boldsymbol{x}'(i)\boldsymbol{x}(i)]^{-1}$ との比較を考える。$\boldsymbol{x}(i)$ は $\boldsymbol{x}$ より i 列を除いたものである。ふたつの正値の対称行列の比較はその行列式の比 $\det[\boldsymbol{x}'(i)\boldsymbol{x}(i)]^{-1}/\det[\boldsymbol{x}'\boldsymbol{x}]^{-1}$ によるのが最も単純でやさしい。$\boldsymbol{x}$ と $\boldsymbol{x}(i)$ は i 列を2乗和と積率の形で含んでいるかいないかの差であるから，この比は1に近く，このことは共分散行列が i 列の除去に対しあまり敏感ではないことを示していると考えてよい。しかしながら，この比は行列 $\boldsymbol{x}$ からの情報のみに基づいており，σ^2 の推定量 s^2 もまた観測値 i の除去によって変動するという事実を反映していない。そこで，s^2 と $s^2(i)$ を導入し，

(3. 15) $\rho\text{Cov.}\equiv\dfrac{\det\{s^2(i)[\boldsymbol{x}'(i)\boldsymbol{x}(i)]^{-1}\}}{\det[s^2(\boldsymbol{x}'\boldsymbol{x})^{-1}]}=\dfrac{s^{2p}(i)}{s^{2p}}\left\{\dfrac{\det[\boldsymbol{x}'(i)\boldsymbol{x}(i)]^{-1}}{\det(\boldsymbol{x}'\boldsymbol{x})^{-1}}\right\}$

を測度として用いることにする。$\det[\boldsymbol{x}'(i)\boldsymbol{x}(i)]=(1-h_i)\det(\boldsymbol{x}'\boldsymbol{x})$ であることから，(3.6) および (3.12) により，ρCov. は

(3. 16) $\rho\text{Cov.}=\dfrac{1}{\left[\dfrac{n-p-1}{n-p}+\dfrac{e_i^{*2}}{n-p}\right]^p(1-h_i)}$

と書き変えることができる。

ρCov. の変動域は，$|e_i^*|\geqq 2$ で $h_i=\dfrac{1}{n}$ のケースと $e_i^*=0$ で $h_i\geqq 2p/n$ のケースの両極端について考えてみると，近似的に $\rho\text{Cov.}\approx 1-\dfrac{3p}{n}$ および

$\rho\mathrm{Cov.} \approx 1+\frac{3p}{n}$ を得る[10]。したがって $|\rho\mathrm{Cov.}-1|$ が $3p/n$ 前後となる観測値は異常値である可能性が大きい。

f) $\mathrm{var}(\hat{y}_i)$

$\hat{y}_i$ の分散の変化の程度にも考慮を払う必要がある。Cov($\boldsymbol{b}$) の場合と同様に観測値 i の除去の前後の分散比をとればよい。$\mathrm{var}(\hat{y}_i)=s^2h_i$, $\mathrm{var}[\hat{y}_i(i)]=\mathrm{var}[\boldsymbol{x}_i{}^r\boldsymbol{b}(i)]=s^2(i)\left[\frac{h_i}{1-h_i}\right]$ であるから,

$$\textbf{(3. 17)}\quad \rho\mathrm{var.}\equiv\frac{s^2(i)}{s^2(1-h_i)}$$

が測度である。これは ρCov. とほぼ同じ変動をすると考えられるが, 巾乗の次数の関係から ρCov. よりも観測値 i の除去にかんして鈍感である。

g) criteria

以上, 異常値を検出するための測度として h_i, e_i*, $\Delta\boldsymbol{b}(i)$, $\Delta^*b_j(i)$, ρCov., $\Delta\hat{y}(i)$, $\Delta^*\hat{y}(i)$, ρvar. を定義した。このうちあらゆる情況のもとで考慮しなければならないのが h_i と e_i* であり, これらについての分布と限界点は各々の項で述べた。$\Delta\boldsymbol{b}(i)$ と $\Delta\hat{y}(i)$ については, これらが標準測度でなく, 観測値の数と行列 $\boldsymbol{x}$ の状態に依存するところが大きいことから一般的な限界点を設定することができないため, その絶対値の大きさの比較を行なうことになる。これに対して, $\Delta^*b_j(i)$, $\Delta^*\hat{y}(i)$ はスケーリングされた測度であるからその限界点を設定することができ, Belsley, Kuh & Welsch (1980) によれば, n を考慮しない 'absolute cutoff'=2.0 である。n を考慮に入れたときの 'size-adjusted cutoff' は $\Delta^*b_j(i)$ が $2/\sqrt{n}$, $\Delta^*\hat{y}(i)$ が $2\sqrt{p/n}$ となる[11]。ρCov. にかんしては size-adjusted cutoff のみをその項で述べておいたが, これは ρCov. には適当な標準誤差スケーリングがないため absolute cutoff が存在しないからである (h_i についても同様のことが言える)。ρvar. についてはほぼ ρCov. と同じであるが, この測度はその項で述べた理由からあまり有用でない。

異常値にかんする本論での定義から，全ての測度のうちのひとつの測度によって，回帰に対して influential であると結論づけされるような観測値は異常値である。しかしこの煩瑣な作業を遂行しても検出されない本来の異常値も存在する場合がある。それは一組のデータに複数個の異常値が含まれている場合で，異常値の数が多くなるほどその傾向は強い。本節でのひとつひとつの観測値に対して検討を試みる方法でも，2つ以上の異常値は十分に検出可能であるが，データに占める異常値の割合が一定限度（10%ぐらいか）を超えた場合には注意を要する。このような場合の問題を次節で論ずる。

注(1) $\boldsymbol{x}$ にはモデルの定数項に対応する全て1よりなる行を含むと考えてよい。

(2) h_i は本来行列の要素であるから h_{ii} と書くべきところを h_i と略記する。また式中の $\boldsymbol{x}_i^r$ は行列 $\boldsymbol{x}$ の第 i 列を示す。

(3) Belsley, Kuh & Welsch (1980), 66 p. $\bar{\boldsymbol{x}}^r$ は独立変数の平均値の列ベクトル。

(4) Rao (1973), 570 p. 邦訳502ページ。

(5) 野口 (1980) を参照せよ。

(6) グラフィックな残差分析の応用については Chaterjee & Price (1977) が詳しい。また基本的な手順に関しては野口 (1980) を参照されたい。

(7) 野口 (1980) 23ページ。

(8) Lund (1975), 473-476 p. また Lund 表は Daniel & Wood (1980), 232-233 p. にも再掲されている。

(9) 本論では‘推定’と‘あてはめ (fit)’とを特に区別していない。

(10) Belsley, Kuh & Welsch (1980), 23 p.

(11) ——, 27-29 p.

4. 異常値集団の検出

複数個の異常値が存在するときの問題を考えてみよう。この場合，異常値の数 (m) がわかっているかどうかで問題の局面は大きく異なる。m が既知であるというケースはあまり存在しないが，この場合は $n-m$ 個の観測値の集団と m 個の観測値の集団との間に，回帰に対する影響に関して有意な差が存在するかどうかを検討すればよい[11]。より一般的な m が未知のケースでは，まず m の決定から始めなくてはならない。これには単一異常値の検出で用いた方法により個々の観測値に対して検討を行うことが賢明な方法であろう。この操

作でmの値を決定することができるが，それと同時にどの観測値が異常値であるかが判明するのであるから，これ以上分析を続ける必要はないのかも知れない。事実，そうなることが多く，その以上の分析は不必要である場合が大部分である。ところが「マスク効果（masking effect)」なるものが存在しているときには，前節の方法を用いても異常値として検出されない事態が起こり得る。

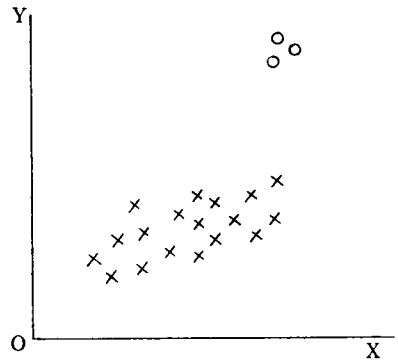

図 2

この効果を図2によって説明しよう。単純化のため独立変数は1個である。図の○印が異常値を示しているとしよう。この場合○が1個しかなければそれを異常値として把えることはやさしい。しかし図のように3個もかたまって存在していると，お互いの「マスク効果」によって各々が×印の集団かけ離れたものであることが隠されてしまうのである。この傾向は○印が多くなるにしたがって強まる。図の場合，回帰直線の勾配は大きくなる。このようなケースの存在の可能性が示唆される場合には，やはり異常値を集団として検出する必要がある。ただし，単一異常値の検出法において各測度の限界点を低く設定する[2]ことによって，マスク効果を取り除くこともある程度可能である。

異常値集団が検出された場合，問題はそれをどう取り扱うかである。集団の大きさがデータに占める割合が小さいのであれば除去すればよく，このような場合には単一異常値の検出方法により全ての異常値が検出できるものである。異常値集団が大きい（データの20%以上）場合の処理は厄介である。まず，異常値集団内部に何らかの規則性（異常値集団が何らかの確率分布を構成しているかどうか）が存在していないかどうかを検討してみる必要がある。規則性がない場合には結局それらを除去するしかなく，全体の標本数が大きければそれで問題はない。標本数が少ないときに20%もの観測値を除去してしまうとデー

タそのものが小さいものになってしまうが，無理に全ての観測値を用いてロバスト推定，等々を行なうのは賢明でない。（異常値に対して）ロバストな統計手法は，異常値の存在が確認し得ないときに行なうべきものであって，ひとたび異常値（集団）の検出がなされたのであれば，これを除去し，正常な（データの生成メカニズムを代表している）観測値によって分析を続けるべきである。尚，本来は何らかの規則性によって支配されている異常値集団であっても，集団の大きさが小さいときにはその規則性が見出されない可能性が大きいことを注意しておく必要がある。

異常値集団に何らかの規則性が確認された場合の処理には一考を要する。規則性をもった異常値集団というのはモデルの想定ミスを反映していることが多く，このような場合，原点に戻ってモデルの構築の段階からやり直すべきであるが，やむを得ずそのまま分析を続けるときにはやはり集団ごと除去してしまうのが良い。(2.2) の slippage model の場合にはダミー変数を使うことにより異常値集団を除去せずにすましても良いであろう。

a) e^*

まず (3.6) で定義された，スチューデント化された残差の拡張から始めよう。全ての観測値を $\boldsymbol{x}_0(n-m\times p)$ と $\boldsymbol{x}_1(m\times p)$ とに分割する。すなわち $(n-m)$ 個の列ベクトル $\boldsymbol{x}_0{}^r$ と m 個の列ベクトル $\boldsymbol{x}_1{}^r$ が得られ，$\boldsymbol{x}_1$ を異常値集団と想定すれば，$\boldsymbol{x}=[\boldsymbol{x}_0 : \boldsymbol{x}_1]$ をもとに得られるスチューデント化された残差の平方和 $\mathrm{SSR}[\boldsymbol{x}_0 : \boldsymbol{x}_1]$ と，$\boldsymbol{x}_1{}^r$ を全てダミー（i 番目の要素のみ 1 で残りはすべて 0 の列ベクトル）でおきかえた場合の平方和 $\mathrm{SSR}[\boldsymbol{x}_0 : m\times\mathrm{dummies}]$ との差は有意でなくてはならない。この差を検討するための測度としては，

$$(4.1)\quad e^*\equiv\frac{\{\mathrm{SSR}[\boldsymbol{x}_0 : \boldsymbol{x}_1]-\mathrm{SSR}[\boldsymbol{x}_0 : m\times\mathrm{dummies}]\}/m}{\mathrm{SSR}[\boldsymbol{x}_0 : \boldsymbol{x}_1]/(n-p-m)}$$

を用いればよく，これは自由度 m および $n-p-m$ の F 分布に従う[3]。

この測度は以下で述べる測度に比べると，計算コストの面では優れているが，マスク効果を取り除く能力は少ない。

b) stepwise method[4]

他の測度について考えてみよう。$\boldsymbol{x}_1^r$ が $\boldsymbol{x}$ の何列目にある列の集合であるかを示す m 個のインデクスの集合を $\boldsymbol{M}$ で表わすことにすれば，(3.7) に対応して

$$(4.2)\quad \Delta\boldsymbol{b}(\boldsymbol{M})\equiv\boldsymbol{b}-\boldsymbol{b}(\boldsymbol{M})$$

となるが，これをスケーリングする適当な標準誤差測度が見あたらないため $\Delta^*b_j(\boldsymbol{M})$ は定義できない。同様に (3.13) も

$$(4.3)\quad \Delta\hat{y}(\boldsymbol{M})\equiv\boldsymbol{x}_j[\boldsymbol{b}-\boldsymbol{b}(\boldsymbol{M})]\qquad j=1, 2, \cdots\cdots, n$$

として定義できるが，$\Delta^*\hat{y}(\boldsymbol{M})$ はやはり定義できない。これらの共分散行列の変化にかんする測度としては，

$$(4.4)\quad \rho\mathrm{Cov}(\boldsymbol{M})\equiv\frac{\det s^2(\boldsymbol{M})[\boldsymbol{x}'(\boldsymbol{M})\boldsymbol{x}(\boldsymbol{M})]^{-1}}{\det s^2(\boldsymbol{x}'\boldsymbol{x})^{-1}}$$

を得る[5]。これらの測度はその絶対値の大小による比較しかできないから，m の値を変え（あるいは $\boldsymbol{M}$ の要素を変更し）ながら段階的な検討（stepwise approach）を試みることになる。例えば，これを (4.3) の $\Delta\hat{y}(\boldsymbol{M})$ を用いておこなうとしよう。出発点は $m=2$ と，$\Delta\hat{y}$ ないし $\Delta^*\hat{y}(i)$ の絶対値の最大と次に大きい観測値をふたつとり，その集合（すなわち $\boldsymbol{x}_1$）を $\boldsymbol{M}_2^{(1)}$ とする。もし (4.3) の上から大きい値ふたつが $\boldsymbol{M}_2^{(1)}$ に含まれているインデクス j を含んでいなければ，他の観測値が現在の観測値にとって代わり $\boldsymbol{M}_2^{(2)}$ が得られる。この手順を (4.3) の上から大きい値ふたつが $\boldsymbol{M}_2^{(k)}$ に含まれるまで k 回くりかえし，$\boldsymbol{M}_2^{(k)}$ が $m=2$ の場合の候補集合となる。次に $m=3$ とし，出発点は $\boldsymbol{M}_2^{(k)}$ をとる（以下同じ）。この手順を $m=4, 5, \cdots\cdots$ とくりかえしていくわけであるが，m のどの値でやめるか，については決定的な規則はなく，$m^\circ\rightarrow m^\circ+1$ と変化した際の，$+1$ に対応する観測値について単一異常値の検出法を試みて，その観測値が異常値でないと判定されたときに m° を終了点とするぐらいがせいぜいである。mの終了点については形式にとらわれず，分析者の判断によった方が良いかも知れない。それでなくても，この stepwise 法

に必要な計算量は膨大で，少なくとも異常値の問題に関する検討が終わった後の分析にかかる計算コストより小さいことの方が少ないであろう。もし「マスク効果」の可能性が存在していないことが明らかであるときには，決して行うべきではない。

c) Wilks' Λ

最後に Wilks の Λ について触れておこう。これは各観測値の中心的位置からの距離を基礎として作られる幾何学的な測度のひとつであり，これまでの測度と異なり回帰の諸結果に対する個別的な検討を与えるものではない。しかし，この幾何学的な測度はデータのグループ化についての知識を持ち合わせていない（すなわち m の値の決定がなし得ない）場合においても，m の値の変動に対する計算コストが低く，また近似的に確率分布を規定できるという利点をもつ。とくに m の値が大きいとき[6]には有用であろう。

これまでの議論から，統計量 h_i と e_i が重要な役割を演じていることは明らかであろう。h_i は $\boldsymbol{x}$ からの情報を伝達し，e_i は $\boldsymbol{y}$ からの情報をもたらしている。この相互関係を観察するための統計量に Rao (1973) によって拡張された Wilks' Λ がある。$\boldsymbol{y}$ と $\boldsymbol{x}$ を合成し，$p+1$ 個の行よりなる併合行列 $\boldsymbol{z}\equiv[\boldsymbol{x}:\boldsymbol{y}]$ を定義しよう。$\boldsymbol{z}$ の各々の列は $p+1$ 次元空間におけるひとつの観測値に対応する。$\boldsymbol{\tau}_1$ を，$\boldsymbol{M}$ に含まれている列に対する要素が 1 で残りが全てゼロである $n\times 1$ ベクトル，また $\boldsymbol{\tau}_2=\boldsymbol{\tau}-\boldsymbol{\tau}_1$ とすれば（$\boldsymbol{\tau}$ は全ての要素が 1 の $n\times 1$ ベクトル），

$$\text{(4. 5)}\quad \Lambda(\boldsymbol{M})=\frac{\det\left[\tilde{\boldsymbol{z}}'\tilde{\boldsymbol{z}}-\left(\frac{1}{m}\right)\tilde{\boldsymbol{z}}'\boldsymbol{\tau}_1\boldsymbol{\tau}_1'\tilde{\boldsymbol{z}}-\left(\frac{1}{n-m}\right)\tilde{\boldsymbol{z}}'\boldsymbol{\tau}_2\boldsymbol{\tau}_2'\tilde{\boldsymbol{z}}\right]}{\det(\tilde{\boldsymbol{z}}'\tilde{\boldsymbol{z}})}$$

が定義される。$\tilde{\boldsymbol{z}}$ は $\boldsymbol{z}$ を $\bar{\boldsymbol{z}}$ によって中心化した行列である。$\tilde{\boldsymbol{p}}\equiv\tilde{\boldsymbol{z}}(\tilde{\boldsymbol{z}}'\tilde{\boldsymbol{z}})^{-1}\tilde{\boldsymbol{z}}'$ とおけば，

$$\text{(4. 6)}\quad \Lambda(\boldsymbol{M})=1-\frac{n}{m(n-m)}(\boldsymbol{\tau}_1'\tilde{\boldsymbol{p}}\boldsymbol{\tau}_1)$$

となる[7]。

$\Lambda(M)$ を用いるには，m の各々の値について (4.6) を計算し，その中で最も小さい値を見つけることである。あるいは $\tilde{z}$ の各列が p 次元正規分布に独立に従うと仮定すれば，近似的に，

$$(4.7)\quad \left(\frac{n-p-1}{p}\right)\left[\frac{1-\Lambda(M)}{\Lambda(M)}\right] \sim F_{p,\ n-p-1}$$

となることが知られている[8]から，これを利用してもよい。

注(1) もし両集団の差を回帰に対しての有意性でなく，分布の位置の差の有意性について検討すればよいのであれば，Kruskal-Wallis 検定や Bell-Doksum 検定に代表されるような 'slippage test' を行なえばよい。slippage test については Conover (1971), 341-347 p. を参照せよ。

(2) 単一異常値の場合，各測度の限界点は通常95%の有意水準をもとに設定されているから，この場合は有意水準を90%程にすればよい。

(3) Gentleman & Wilk (1975), 387-411 p.

(4) ここでいう stepwise method は，いわゆる 'stepwise regression' とは異なる。

(5) ρ var.(M) についてはあまり重要でないので省略した。

(6) Belsley, Kuh & Welsch (1980) によれば，$m \leqslant 20$ まで計算可能で，それ以上は Mahalanobis' D 等によればよい。38 p.

(7) ——, 67 p.

(8) Rao (1973) 570 p. 邦訳502ページ。

文 献

[1] Barnett, V. & T. Lewis (1978), *Outliers in Statistical Data,* John Wiley & Sons.

[2] Belsley, D. A., E. Kuh & R. E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley & Sons.

[3] Chatterjee, S. & B. Price (1977), *Regression Analysis by Example,* John Wiley & Sons (佐和隆光，加納悟訳「回帰分析の実際」新曜社，1981).

[4] Conover, W. J. (1971), *Practical Nonparametric Statistics,* John Wiley & Sons.

[5] Daniel, C. & F. S. Wood (1980), *Fitting Equations To Data: Computer Analysis of Multifactor Data* (2nd. ed.), John Wiley & Sons.

[6] David, H. A. (1979), *Robust Estimation in the Presence of Outliers,* a contribution to *Robustness In Statistics* (edited by Launer, R. L. & G. N. Wilkinson, Academic Press).

[7] Dixon, W.J. (1953), *Processing data for outliers,* Biometrics (9), pp. 74–89.

[8] Gentleman, A. F. & M. B. Wilk (1975), *Detecting Outliers II Supplementing the Direct Analysis of Residuals,* Biometrica (62), pp. 387–411.

[9] Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* John Wiley & Sons (丘本正・磯具恭史訳「統計的多変量データ解析」日科技連出版社, 1979)。

[10] Hawkins, D. M. (1980), *Identification of Outliers,* Chapman & Hall.

[11] Lund, R. E. (1975), *Tables for an Approximate Test for Outliers in Linear Models,* Technometrics (17), pp. 473–476.

[12] Mosteller, F. & J. W. Tukey (1977), *Data Analysis and Regression,* Addison-Welsley: Reading, Mass.

[13] 野口和也 (1980),「回帰残差の検討」早稲田大学大学院経済学研究年報 (第20号), pp. 13–27.

[14] Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2 nd. ed.), John Wiley & Sons (奥野忠一, 他訳「統計的推測とその応用」東京図書, 1977)。

1981.9.30 脱稿

(後期課程第2年度生・統計学保田順三郎研究指導)